A PC Classifier of Clinical Text Documents: Advanced Information Retrieval Technology Transfer

David B. Aronow, M.D., M.P.H., Avinoam Shmueli, BS
Center for Intelligent Information Retrieval
University of Massachusetts
Amherst, MA
{aronow, shmueli}@cs.umass.edu

The Center for Intelligent Information Retrieval (CIIR) and Harvard Pilgrim Health Plan (HPHC) have collaborated on a research project to determine the extent to which automated information retrieval systems can reduce the burden of manual chart review for quality measurement in health care organizations. This technology is being transferred to HPHC via the creation of a PC-based application for the classification of clinical text documents.

The HPHC Inquery Classifier is a Windows 95™ application, comprised of a Dynamic Link Library encapsulating the Inquery Information Retrieval Engine, and a graphical interface written in Visual Basic™.

The Classifier calls Inquery, an information retrieval system based on Bayesian inference networks. Inquery's relevance feedback and logistic regression features create the classification profile query and sort documents into relevancy bins. The sort is based on evidence found by Inquery within the text of a representative sample of pre-scored training documents.

There are five steps to classification:

1. prepare raw data
2. build indices for inference network
3. score training documents
4. build classification profiles and test their effectiveness
5. classify new documents into Positive, Uncertain and Negative Bins

The user then focuses manual review on the Uncertain Bin documents only, as shown in the Figure. Initial experiments classifying automated medical record encounter notes for asthmatic children produced average precisions in the range of 80% with a reduction in manual chart review projected up to 80%.

This material is based on work supported by NRaD Contract Number N66001-94-D-6054. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect those of the sponsor.